nature computational science

ARTICLES

https://doi.org/10.1038/s43588-022-00266-5

Check for updates

Cell clustering for spatial transcriptomics data with graph neural networks

Jiachen Li[1,2], Siheng Chen[3,4], Xiaoyong Pan[1,2], Ye Yuan[1,2] ✉ and Hong-Bin Shen[1,2] ✉

Spatial transcriptomics data can provide high-throughput gene expression profiling and the spatial structure of tissues simultaneously. Most studies have relied on only the gene expression information but cannot utilize the spatial information efficiently. Taking advantage of spatial transcriptomics and graph neural networks, we introduce cell clustering for spatial transcriptomics data with graph neural networks, an unsupervised cell clustering method based on graph convolutional networks to improve ab initio cell clustering and discovery of cell subtypes based on curated cell category annotation. On the basis of its application to five in vitro and in vivo spatial datasets, we show that cell clustering for spatial transcriptomics outperforms other spatial clustering approaches on spatial transcriptomics datasets and can clearly identify all four cell cycle phases from multiplexed error-robust fluorescence in situ hybridization data of cultured cells. From enhanced sequential fluorescence in situ hybridization data of brain, cell clustering for spatial transcriptomics finds functional cell subtypes with different micro-environments, which are all validated experimentally, inspiring biological hypotheses about the underlying interactions among the cell state, cell type and micro-environment.

A number of spatial transcriptomics technologies have been developed to achieve high-throughput gene expression profiling and the spatial structure of tissues simultaneously. Some of these achieve single-cell resolution by using in situ hybridization, such as cyclic ouroboros single-molecule fluorescent in situ hybridization[1], multiplexed error-robust fluorescence in situ hybridization (MERFISH)[2–5], sequential fluorescence in situ hybridization (seqFISH)[6,7], enhanced seqFISH (seqFISH+)[5] and spatially resolved transcript amplicon readout mapping[8], which can quantify the RNA transcripts of genes and their locations in the sample. Integrated with image analysis, fluorescence in situ hybridization (FISH) enables single cell-resolution high-throughput gene expression quantification and spatial location recording. Alternative approaches include RNA sequencing (RNA-Seq)-based technologies such as spatial transcriptomics (ST)[9], Slide-seq[10], Slide-seqV2[11], laser capture microdissection coupled with full-length messenger RNA sequencing (LCM-seq)[12], etc. While these methods lead to whole-transcriptome profiling, most cannot provide single-cell resolution. Some approaches such as BayesSpace[13] and stereoscope[14] have been developed to enhance the resolution of spatial transcriptomic data based on probabilistic models.

An essential question regarding single-cell gene expression data is the cell state or type identification, which is one of the key steps in the data processing pipeline, including lineage[15], cell cycle[5], cell–cell interaction analysis[16,17], etc. Several clustering approaches have been developed for single-cell RNA-Seq data, mainly being based on clustering of a low-dimension representation of the gene expression of single cells[18–21]. Most spatial data studies also rely on such strategies. For the MERFISH dataset[5], graph-based Louvain community detection[22,23] is applied to the top principal components of the gene expression of single cells[24,25]. Integration of single-cell RNA-Seq (scRNA-Seq) was also adopted in previous work[26].

For spatial data, these expression-based methods cannot make full use of spatial location information, which is often coupled with cell identities. In vitro cultured cells in the same cell cycle phase are more likely to reside together[5], and certain cell types of in vivo tissue are known to be spatially proximal to themselves or to specific cell types[27]. The spatial structure could thus be used as an informative feature to improve cell clustering in recent studies. Giotto[28] is a package designed for processing spatial gene expression data. stLearn[29] firstly utilizes the standard Louvain clustering procedure as used in scRNA-Seq analysis to obtain a *k*-nearest neighbour graph. Next, the initial cluster is split into subclusters if its spots are spatially separated. SmfishHmrf[30] is another spatial clustering method that starts from the support vector machine classifier trained using scRNA-Seq data as mentioned above. It then updates the cell clustering according to the principle that neighbouring cells with the same identity have higher scores. BayesSpace[13] is a Bayesian statistical clustering method designed for only ST data which encourages neighbouring spots to belong to the same cluster. SpaGCN[31] utilizes a graph convolutional network (GCN) to integrate gene expression with spatial location and histology in ST. Spatial embedded deep representation (SEDR)[32] uses a deep autoencoder to map the gene latent representation to a low-dimensional space. Spatial transcriptome-based cell-type clustering (STEEL)[33] is a manifold learning-based algorithm for cell type identification from the spatial transcriptome. Most of these existing spatial clustering approaches assume that the same cell group is spatially close to each other and do not take into consideration the whole complex global cell interactions across the tissue sample. Much work still needs to be carried out on this promising spatial representation.

Here, we develop a cell clustering method called cell clustering for spatial transcriptomics data (CCST), based on GCNs, which can combine the gene expression and complex global spatial information of single cells from spatial gene expression data. A few years ago, the GCN[34] was introduced to handle non-Euclidean structural data by encoding it as a graph with an adjacency matrix representing the relationships among variables and a node feature matrix

[1]Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China. [2]Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai, China. [3]Cooperative Medianet Innovation Center (CMIC), Shanghai Jiao Tong University, Shanghai, China. [4]Shanghai Artificial Intelligence Laboratory, Shanghai, China. ✉e-mail: yuanye_auto@sjtu.edu.cn; hbshen@sjtu.edu.cn

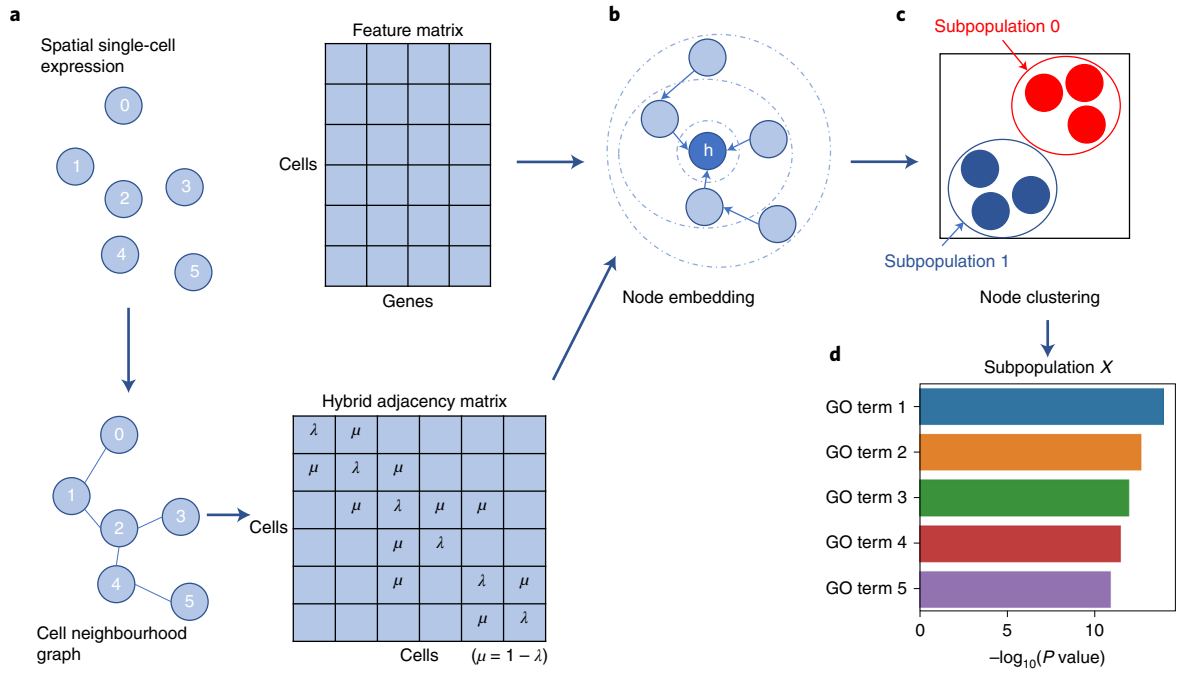


Fig. 1 | CCST workflow for cell subpopulation discovery. a, The data preprocessing, including the generation of a cell neighbourhood graph according to their spatial location, the preprocessing of gene expression information to obtain the cell feature matrix and the construction of the hybrid adjacency matrix using the hyperparameter λ. **b**, The application of DGI for node embedding with spatial information and PCA for further dimension reduction. h means the node representation in the hidden space. Taking the node in dark blue for example, features can be aggregated within the first-order neighborhoods in each graph convolutional layer. As a result, the receptive field (dash-dotted rings) extends with the increase in number of graph convolutional layers. **c**, Node clustering for cell subpopulation discovery. **d**, DE analysis with the Mann–Whitney *U* test and GO analysis.

representing the variable observations. For the cell clustering of spatial data, we first convert the data to a graph where a node represents a cell with its gene expression profile as attributes and an edge represents the neighbourhood relationship between cells. Next, a series of GCN layers is used to transfer the graph and gene expression information as cell node embedding vectors, while the graph is corrupted to generate negative embeddings. By learning the discrimination task, the neural network is trained to encode the cell embedding from spatial gene expression data, which is then used for cell clustering.

The CCST method is tested on both FISH-based single-cell transcriptomics and spot-based ST. CCST is also tested on both in vitro and in vivo spatial datasets for the tasks of ab initio cell clustering and cell subtype discovery based on a manually curated cell category annotation. Our experimental results suggest that, in comparison with prior methods, CCST can improve the ab initio cell clustering in the MERFISH dataset[5] by clearly recognizing cell groups in all four cell cycle phases among cultured cells of the same cell type. CCST can also be used to find cell subtypes and their interactions, providing biological insights from seqFISH+ datasets of mouse olfactory bulb (OB) and cortex tissues[35]. In addition, to illustrate its performance in comparison with recently developed methods, CCST is evaluated on two ST datasets and achieves better clustering results. All of these results indicate that CCST can provide informative clues to improve understanding of the cell identity and interactions as well as the spatial organization of tissues and organs.

Results

The CCST framework. We extended the unsupervised node embedding method Deep Graph Infomax (DGI)[36] and developed CCST to discover cell subpopulations from spatial single-cell expression data. As shown in Fig. 1, with both the single-cell location and gene expression information as inputs, CCST firstly encodes the spatial data into two matrices. One is a hybrid adjacency matrix based on cell neighbourhoods, where a hyperparameter λ is used to balance intracellular (gene) and extracellular (spatial) information (Methods), while the other one is the single-cell gene expression profile matrix. Both matrices are fed into the DGI network to calculate an embedding vector for each cell. DGI employs a series of GCN layers, which enables it to integrate both graph (cell location) and node attributes (gene expression) as node (single cell) embedding vectors. The edges in the graph are also permuted to generate negative node embedding vectors that do not have any spatial structure information. After being trained to discriminate the two embedding types, CCST learns to encode a cell node embedding that contains both the spatial structural and gene expression information. After dimension reduction by principal component analysis (PCA), the *k*-means++ algorithm is used for node clustering to identify cell groups or subpopulations.

Applying CCST to spatial gene expression data. While a number of spatial gene expression datasets have been created, here we focus on three FISH-based datasets that contain thousands of genes with single-cell resolution. The first is the MERFISH data[5] obtained from in vitro cultured U-2 OS cells with different cell cycle phases. As mentioned by the authors of the MERFISH paper, obvious spatial structures of cell cycle phase were discovered within this cell population, so it represents an ideal spatial dataset to test clustered cell groups since cell cycle can be used as ground truth here. The other two are seqFISH+ datasets[35], which include several in vivo cell types and so can be used to explore potential cell subpopulations with complex biological molecular and spatial features. See Methods section for dataset and preprocessing details.

Although CCST is designed to identify cell subtypes and single-cell interactions, it is also applied to two more ST datasets here, to test its generalizability and extend its potential application scope.

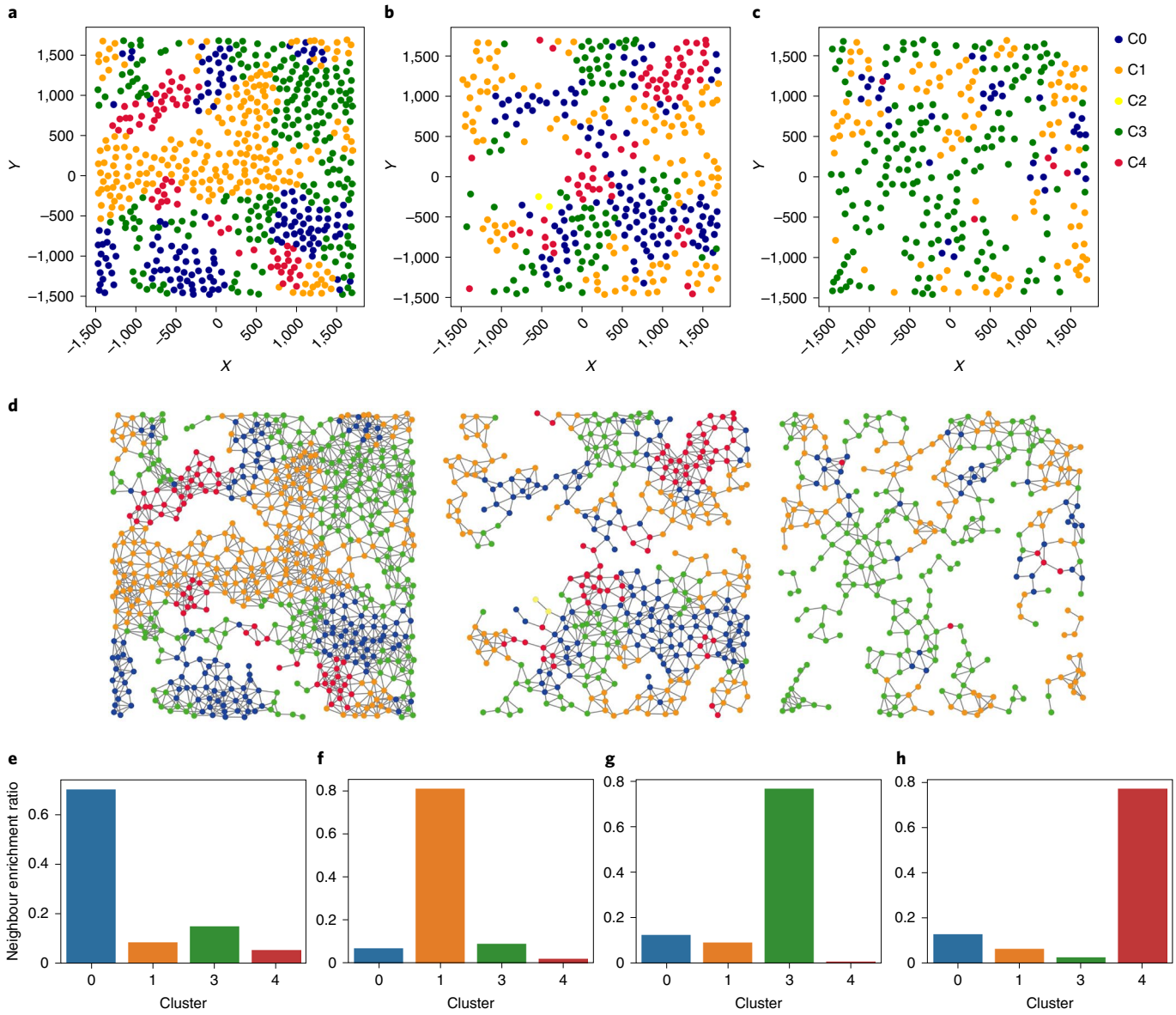


Fig. 2 | Spatial distribution of cells in different clustered groups in MERFISH dataset. a–c, The spatial distributions of cells from replicate batches 1 (**a**), 2 (**b**) and 3 (**c**). *X* and *Y* are the coordinate axes of spatial location. The colour assignment to each cluster shown in the legend to panel **c** applies to all panels. **d**, The constructed graphs, omitting nodes without neighbours. Note that CCST merges all the cells from different batches into one whole graph by constructing a block-diagonal adjacency matrix. **e–g**, The neighbour enrichment ratios for C0 (**e**), C1 (**f**), C3 (**g**) and C4 (**h**).

These two ST datasets are the human dorsolateral prefrontal cortex (DLPFC)[37] dataset and the 10× Visium spatial transcriptomics dataset of human breast cancer. Multiple metrics are used to evaluate the performance of CCST and other approaches.

Identifying spatial heterogeneity from the MERFISH dataset. We first assess the ability of CCST to cluster cells using the cultured U-2 OS MERFISH dataset. During the data preprocessing, the cells from all batches were merged by constructing a block-diagonal adjacency matrix (Supplementary Fig. 1). CCST was then trained with the normalized gene expression matrix and the hybrid adjacency matrix obtained from the spatial structure to generate the embedding vector. To further reduce the feature dimension, PCA was performed, and the top 30 principal components selected for *k*-means clustering with a *k* value of 5, as suggested in the MERFISH study[5]. Given the fact that cluster 2 (C2) includes only two cells, the analysis below focuses on the other four groups.

Figure 2a–c shows the spatial distribution of grouped cells obtained by CCST on all three replicates. Cells in clusters C0 to C4 are shown by points of different colours, located at the centre of each cell. To make full use of the dataset, we encode all three replicates with just one adjacency matrix, where cells are only connected within each batch such that the matrix is a block-diagonal matrix composed of three adjacency submatrices (Methods and Supplementary Fig. 1). To further investigate the neighbourhood spatial structure of cells assigned to different groups, the neighbour enrichment ratios for C0, C1, C3 and C4 are shown in Fig. 2e–h. For all cells in a certain group, we first collect their neighbouring cells according to the initial adjacency matrix. Next, we count how many of the neighbours are assigned to each group, and calculate their proportions as the neighbour enrichment ratios. These ratios clearly show that cells tend to spatially neighbour those of the same group, which is similar to the conclusion in MERFISH literature. As discussed in the next section, Gene Ontology (GO) term analysis

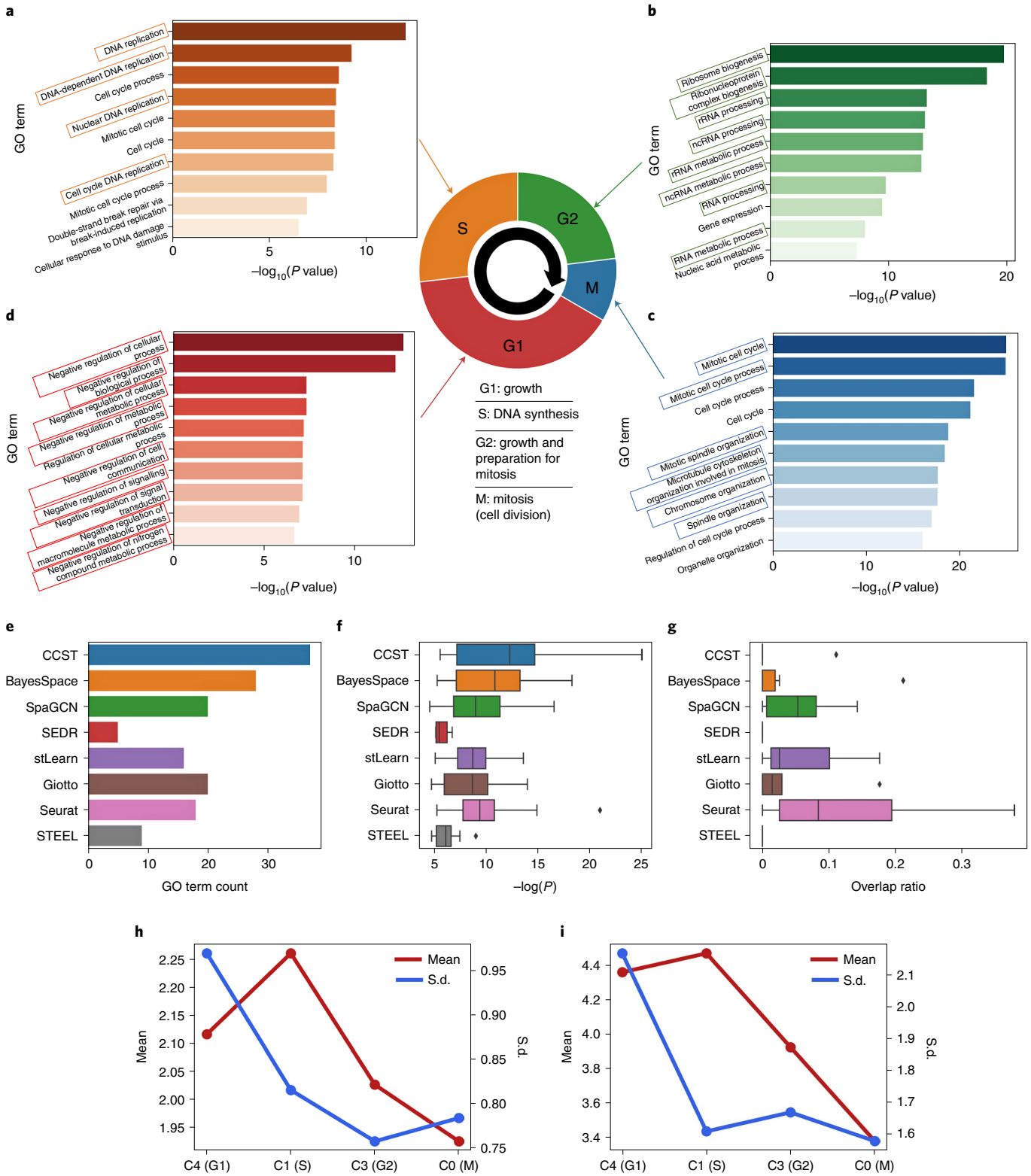


Fig. 3 | Cell cycle phase identification. a–d, The top GO terms for the clustered cell groups C1 (**a**), C3 (**b**), C0 (**c**) and C4 (**d**), corresponding to S, G2, M and G1 phase, respectively. Key GO terms are boxed. **e–g**, Comparison of GO term results for CCST versus prior methods in terms of the total number of GO terms (**e**), significance of key GO terms (**f**) and overlap ratio of GO terms (**g**) among different cell groups. **f** shows a box plot of statistical significance associated with the key GO terms obtained by CCST ($n=37$), BayesSpace ($n=28$), SpaGCN ($n=20$), SEDR ($n=5$), stLearn ($n=16$), Giotto ($n=20$), Seurat ($n=18$) and STEEL ($n=9$). **g** shows a box plot of the overlap ratio between the GO terms from each pair of clusters obtained by CCST ($n=6$), BayesSpace ($n=6$), SpaGCN ($n=6$), SEDR ($n=6$), stLearn ($n=1$), Giotto ($n=3$), Seurat ($n=6$) and STEEL ($n=3$). In **f** and **g**, the left, central and right hinges correspond to the first quartile, median value and third quartile, respectively. The whiskers extend to 1.5× the interquartile range of the distribution from the hinge. Data beyond the end of the whiskers are determined to be outliers and plotted individually as diamonds. **h**,**i**, The mean and s.d. of *CDT1* (**h**) and *CDC6* (**i**). The number of cells included in each cluster is 304 (C0), 448 (C1), 472 (C3) and 142 (C4).

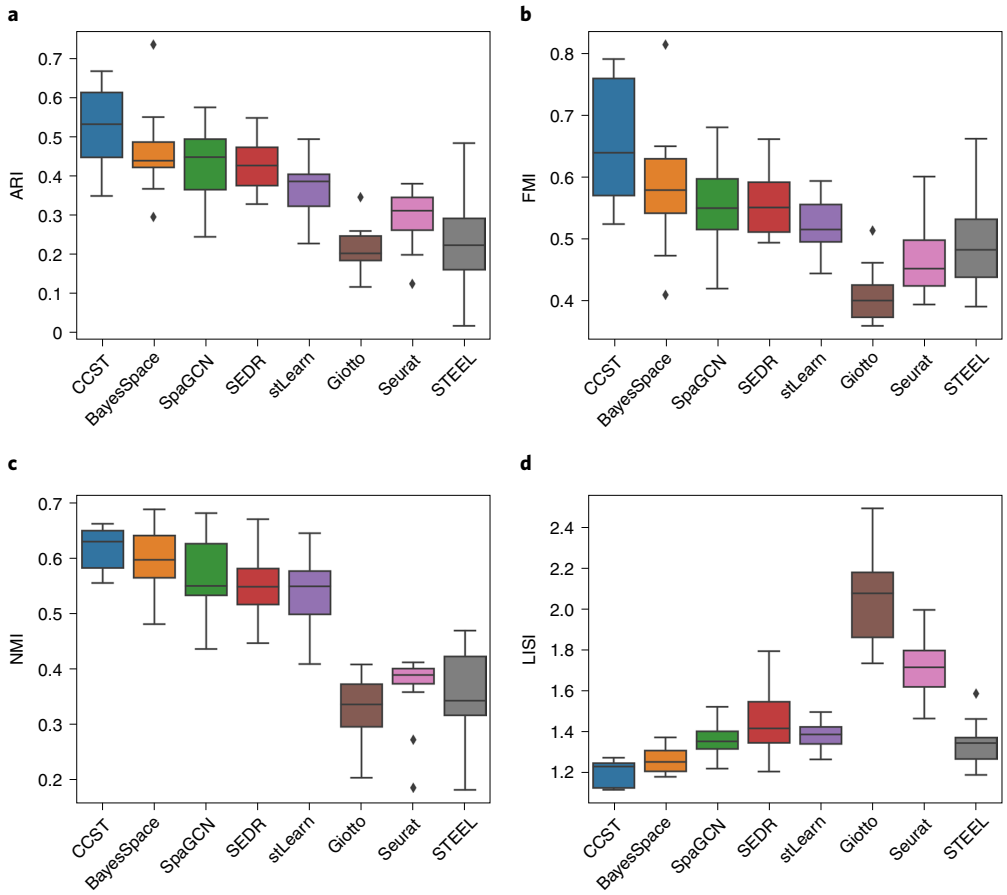


Fig. 4 | Performance of CCST on two annotated datasets. a–d, Box plots of ARI (**a**), FMI (**b**), NMI (**c**) and LISI (**d**) for CCST and prior methods on 12 samples from the DLPFC dataset ($n = 12$ biologically independent samples). The lower, central and upper hinges correspond to the first quartile, median value and third quartile, respectively. The whiskers extend to 1.5× the interquartile range of the distribution from the hinge. Data beyond the end of the whiskers are considered as 'outliers' and plotted individually as diamonds.

suggests that each cluster corresponds to one cell cycle phase exclusively (M for C0, S for C1, G2 for C3, G1 for C4). note also that C0 (M) is spatially proximal to C3 (G2), as is C1 (S) to C3 (G2) and C4 (G1) to C0 (M), which indicates that cells of adjacent phases co-locate with each other, as well. This could be explained by the fact that spatially proximal cells may have divided from the same mother cell.

Identifying cell cycle phases. We next perform differential expression (DE) analysis to verify the different biological functions of each clustered cell group. Here, the Mann–Whitney *U* test is used to find highly expressed DE genes in each cell group compared with all other groups. Then GO term enrichment analysis is done using the top 200 significantly DE genes with the whole MERFISH gene list as the background gene set, with the result sorted in descending order by −log(*P* value) (Fig. 3). These results indicate that CCST can clearly identify all four cell cycle phases. The significantly highly expressed genes in C1 are mostly related to GO terms for DNA replication, indicating that C1 refers to cells in the S phase, the stage during which DNA is replicated. The DE genes in C3 are mostly related to GO terms for RNA processing, indicating that cells in C3 are mainly in G2 phase, when macromolecules for multiplication and cell growth are produced, preparing for the next M stage. The top GO terms in C0 correspond to the mitotic cell cycle process, indicating that cells in C0 are in M (mitosis) phase, when cells give birth to new progeny cells. C4 is enriched with GO terms corresponding to the negative regulation of various processes, indicating that these cells are in G1 phase, resting in preparation for the next cell cycle. Although the G1 phase is very complicated, including a variety of biological processes[38], the top DE genes further confirm the CCST predictions (Supplementary Data 1). The most highly DE genes in C4 are *MALAT1* and *ABI2*, both of which are related to the cell cycle and play important roles in the G1-to-S phase transition[39–41]. In addition, *CDT1* and *CDC6* are essential for the initiation of DNA replication and are well-known marker genes for the cell cycle stage. The mean and variance of *CDT1* and *CDC6* (Fig. 3h,i) also support our predictions according to the findings in previous studies[42,43] that *CDT1* increases from a very low level in G1 and starts to decrease after entering the S stage, with its expression varying most in G1. These results indicate that CCST can identify all four cell cycle phases and that C1, C3, C0 and C4 correspond to phase S, G2, M and G1, respectively.

We compare CCST with six recently developed approaches for spatial expression analysis (Giotto[28], stLearn[29], SEDR[32], BayesSpace[13], SpaGCN[31] and STEEL[33]), one scRNA-Seq method (Seurat[21]) and the result in the MERFISH paper[5] where only gene expression is utilized. To enable an overall quantitative comparison among the methods, firstly, by defining key GO terms for each cell cycle (see Supplementary Section 5 for details), we show that CCST finds the highest number of key GO terms (Fig. 3e) with the most statistical significance (Fig. 3f). We then analyse the overlap between the GO terms of the different cell groups for each method. A lower overlap ratio indicates better performance (see Supplementary Section 13 for details), and Fig. 3g shows that CCST achieves very

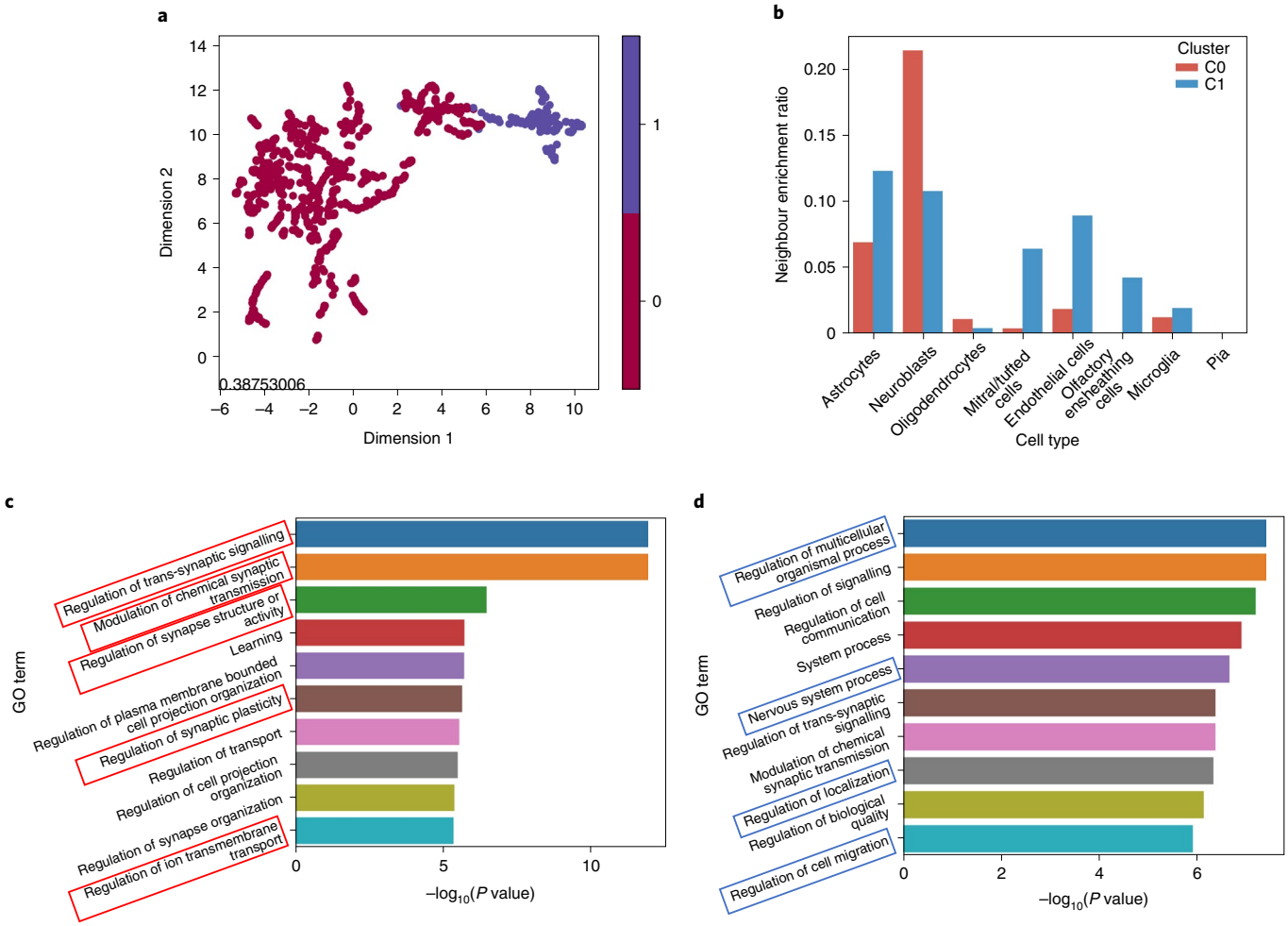


Fig. 5 | Identifying cell subgroups in interneuron cells of the seqFISH+ mouse OB dataset. a, The two-dimensional uniform manifold approximation and projection clustering result with the silhouette coefficient. **b**, Bar plot of the neighbour enrichment ratios for two subgroups. Here, the cells of interneurons are excluded from the histogram to enable a clearer demonstration of the difference in the distribution. **c**,**d**, The significant GO terms based on the top 200 highly DE genes for C0 (**c**) and C1 (**d**). Boxed GO terms indicate that these two subpopulations are functionally different.

promising results. SEDR and STEEL also show low overlap rates, but this is because these two methods can only generate much less significant GO terms. Details about the comparison can be found in Supplementary Section 5. The results suggest that the CCST clusters can be best explained by the cell cycle according to the GO terms, and show the clearest spatial neighbourhood structure[5].

Comparison of CCST with prior methods on ST datasets. The first ST dataset we used was the Lieber Institute for Brain Development human dorsolateral prefrontal cortex (DLPFC)[37] dataset including the 10× Genomics Visium spatial transcriptomics and manually annotated layers. There are 12 samples in the DLPFC data, each comprising up to six cortical layers and the white matter. To enable a comprehensive comparison, we compare CCST with seven recently proposed methods. For a fair comparison using the same input, we only feed the location of each spot and the gene expression matrix into these models. To measure the consistency between the clustering labels and reference labels, the adjusted Rand index (ARI), Fowlkes–Mallows index (FMI) and normalized mutual information (NMI) are employed to compare the performance of the different clustering algorithms (the higher the better) (Fig. 4a–c). Spots in the same biological layer in the brain should be spatially close to each other but separated between different layers. To quantify this property, the local inverse Simpson's index (LISI)[44] is introduced[32] as a metric to access the local diversity of cells (Fig. 4d). A lower LISI value indicates that clusters are better spatially separated. The annotation and clustering results obtained by each method on slice 151676 of the DLPFC dataset are shown in Extended Data Fig. 1. As can be seen, CCST is the closest to the annotated layer segmentation numerically and can find the hierarchical structure of the layers with significantly smoother boundaries, as shown by the results for all of the methods on all 12 samples in Supplementary Fig. 19.

The CCST method is also tested on one more ST dataset, the 10× Visium spatial transcriptomic data of human breast cancer with the manual annotation provided in SEDR[32], which has 20 regions and 4 main morphotypes: ductal carcinoma in situ/lobular carcinoma in situ (DCIS/LCIS), healthy tissue (Healthy), invasive ductal carcinoma (IDC) and tumour surrounding regions with low features of malignancy (Tumor edge). Here we only compare ARI, FMI and NMI rather than LISI because tumor tissues are highly heterogeneous. The annotation and cluster results obtained using each method are shown in Extended Data Fig. 2. Again, the CCST cluster shows a smoother boundary while the clusters obtained using the other compared methods are more fragmented with spot-level noise. Quantitatively, CCST achieves the highest ARI and NMI values among all the methods, and a competitive NMI value when compared with SpaGCN and SEDR.

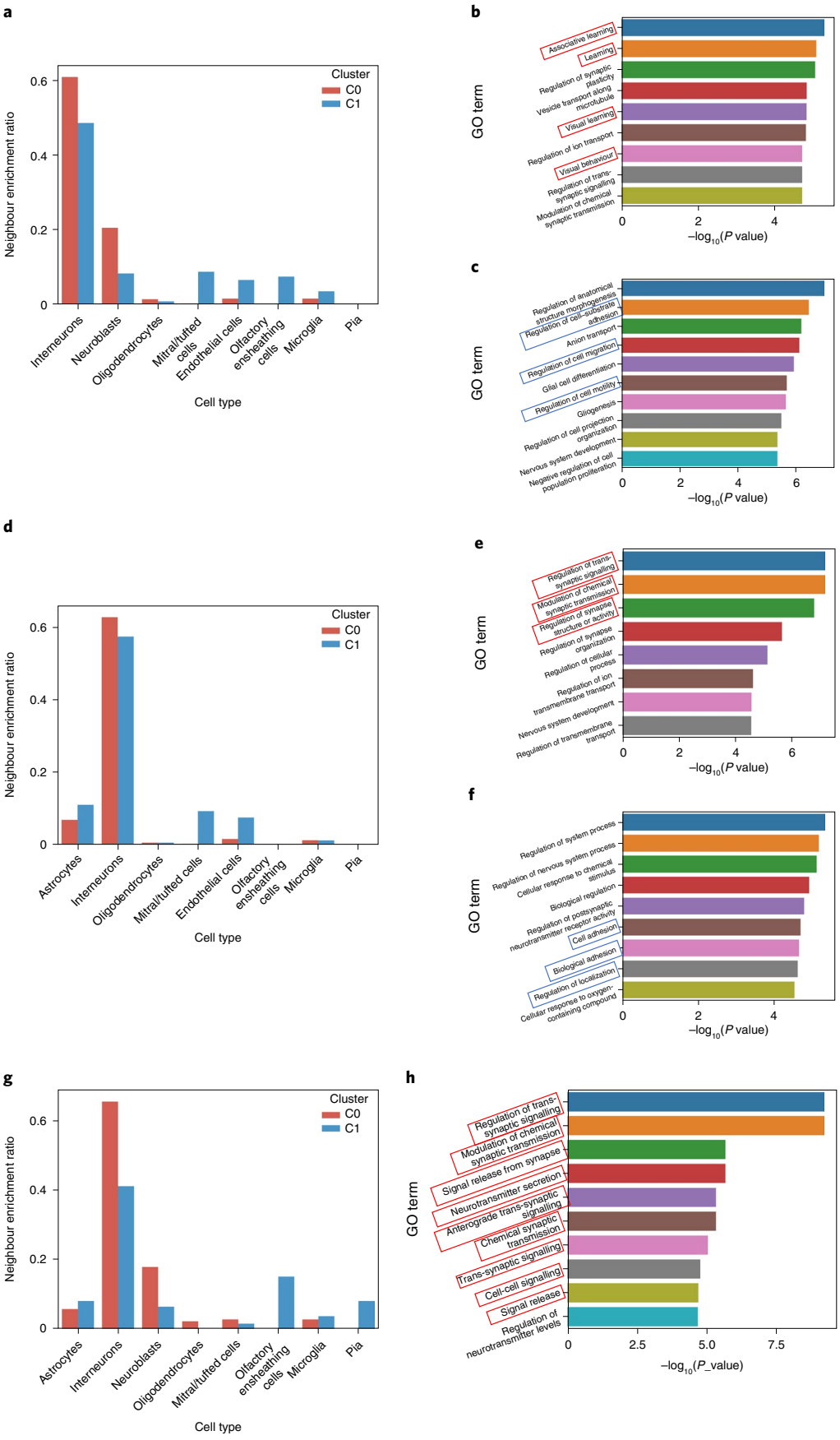

a
Neighbour enrichment ratio
Cluster
C0
C1
Interneurons
Neuroblasts
Oligodendrocytes
Mitral/tufted cells
Endothelial cells
Olfactory ensheathing cells
Microglia
Pia
Cell type
b
GO term
Associative learning
Learning
Regulation of synaptic plasticity
Vesicle transport along microtubule
Visual learning
Regulation of ion transport
Visual behaviour
Regulation of trans-synaptic signalling
Modulation of chemical synaptic transmission
–log10(P value)
c
GO term
Regulation of anatomical structure morphogenesis
Regulation of cell–substrate adhesion
Anion transport
Regulation of cell migration
Glial cell differentiation
Regulation of cell motility
Gliogenesis
Regulation of cell projection organization
Nervous system development
Negative regulation of cell population proliferation
–log10(P value)
d
Neighbour enrichment ratio
Cluster
C0
C1
Astrocytes
Interneurons
Oligodendrocytes
Mitral/tufted cells
Endothelial cells
Olfactory ensheathing cells
Microglia
Pia
Cell type
e
GO term
Regulation of trans-synaptic signalling
Modulation of chemical synaptic transmission
Regulation of synapse structure or activity
Regulation of synapse organization
Regulation of cellular process
Regulation of ion transmembrane transport
Nervous system development
Regulation of transmembrane transport
–log10(P value)
f
GO term
Regulation of system process
Regulation of nervous system process
Cellular response to chemical stimulus
Biological regulation
Regulation of postsynaptic neurotransmitter receptor activity
Cell adhesion
Biological adhesion
Regulation of localization
Cellular response to oxygen-containing compound
–log10(P value)
g
Neighbour enrichment ratio
Cluster
C0
C1
Astrocytes
Interneurons
Neuroblasts
Oligodendrocytes
Mitral/tufted cells
Olfactory ensheathing cells
Microglia
Pia
Cell type
h
GO term
Regulation of trans-synaptic signalling
Modulation of chemical synaptic transmission
Signal release from synapse
Neurotransmitter secretion
Anterograde trans-synaptic signalling
Chemical synaptic transmission
Trans-synaptic signalling
Cell–cell signalling
Signal release
Regulation of neurotransmitter levels
–log10(P_value)

Fig. 6 | Neighbour enrichment ratios and GO term analysis for each cell subtype of astrocytes, endothelial cells, and neural stem cells of the seqFISH+ mouse OB dataset. a–**h**, Neighbour enrichment ratio (**a**,**d**,**g**) and GO term analysis for C0 (**b**,**e**,**h**) and C1 (**c**,**f**) for astrocytes (**a**–**c**), neuroblasts (**d**–**f**) and endothelial cells (**g**,**h**) where there is no significant GO term for C1, showing the difference in neighbourhood enrichment of two subtypes or the difference in GO terms between them. Boxed GO terms indicate that the two subpopulations are functionally different.

Identifying cell subtypes on the seqFISH+ dataset. In addition to ab initio discovery of cell groups, we next show that CCST can also be used to identify cell subtypes and interactions from a manually curated cell type annotation based on prior biological knowledge. For this, we firstly select the seqFISH+ dataset from mouse OB. We apply the CCST method to all 11 cell types to identify subpopulations within each annotated cell type. With the same hyperparameter settings as used for the MERFISH dataset, the embedding vector generated by GCN is fed into PCA, and the top 30 principal components are utilized to divide each annotated cell type group into two clusters to discover potential cell subtypes.

We first analyse the cell subtype result for interneuron cells (Fig. 5). On the basis of the spatial embedding, the annotated interneuron cells can be clearly divided into two subgroups in the reduced two-dimension uniform manifold approximation and projection space (Fig. 5a). The barplots of the neighbour enrichment ratios for the two subgroups (Fig. 5b) indicate that the two subsets of cells have very different micro-environments. Specifically, cells in C1 tend to be spatially proximal to mitral/tufted cells, endothelial and olfactory ensheathing cells. We then performed GO term analysis based on the top 200 differentially highly expressed genes in C0 (Fig. 5c) and C1 (Fig. 5d). For C0, the top enriched GO terms are relevant to neural functions such as trans-synaptic signalling. In contrast, the top GO terms for C1 are not so related to neural functions, instead including regulation of localization and cell migration. In addition, the most significantly highly expressed gene in C1 is *NRSN1* ($P = 1.67 \times 10^{-35}$), which is important for neural organelle transport, nerve growth and neurite extension[45]. Such results indicate that the interneuron cells can be divided into two subgroups: a functionally mature neural cell group that can communicate with other neural cells, and a group of cells that is still in development, including localization and migration, which are interacting with neighbouring cells such as mitral/tufted or endothelial cells. Interestingly, this discovery of cell subtypes and their interactions is validated by the recent findings that, in a subclass of interneuron, GABA (γ-aminobutyric acid)-ergic interneuron migration can be regulated via embryonic forebrain endothelial cells[46] and that the partial loss of GABA released from endothelial cells can impair the long-distance migration and localization of interneurons during embryogenesis.

Distinct micro-environment settings of two cell subgroups within one annotated cell type are also found for astrocyte, neuroblast and endothelial cells (Fig. 6). The C0 subgroup of astrocyte cells are more spatially proximal to interneurons, neuroblasts, etc (Fig. 6a), and its DE genes are mainly enriched in GO terms related to visual and learning functions, while the C1 subgroup of astrocytes has GO terms related to cell migration (Fig. 6b,c). For neuroblast cells, we find a pattern similar to that of interneuron cells. The C1 cluster of neuroblasts are more spatially close to mitral/tufted or endothelial cells compared with C0 (Fig. 6d). The top GO terms for C0 are relevant to neural functions, such as regulation of trans-synaptic signalling, indicating that C0 are functional mature neural cells, while C1 is unmatured and related to cell adhesion according to its GO terms (Fig. 6e,f). Moreover, the most significantly high expressed gene in C1 is *EOMES* ($P = 5.57 \times 10^{-39}$), which is essential for the central nervous system in vertebrates[47]. The sub-neuroblast discovery is also supported by a very recent finding[48] of direct contacts between endothelial cells and some neuroblasts. For endothelial cells, we also find two cell subtypes with different micro-environments. Specifically, we find GO terms associated with synaptic signalling and cell–cell signalling for the C0 group (Fig. 6g,h). Further neighbour enrichment analysis for all four cell types finds more complex cell interactions with cell subtype resolution (Supplementary Fig. 27).

For comparison, we also perform CCST with $\lambda = 1$, where no spatial structure information is taken into consideration. However, no significant GO term is found for all cell subtype groups, indicating that spatial information is essential to discover cell subtypes. We also apply such experiments and analysis to the seqFISH+ mouse cortex dataset. The results are shown in Supplementary Figs. 28–29.

Discussion

To make full use of spatial and gene expression level information, here we introduce CCST, which uses unsupervised GCNs to learn a cell embedding representation based on graphs extracted from spatial transcriptomics data. Unlike the assumption made in most existing approaches that the same cell group is spatially close to each other, CCST takes into consideration all the complex global cell interactions across the tissue sample. The experimental results suggest that CCST is a promising algorithm to help improve understanding of cell identity, interaction and spatial organization from spatial data.

Despite the attempt shown in Supplementary Fig. 25, how to select the cluster number accurately remains a limitation when applying CCST to new datasets without prior knowledge, which is also a general problem for unsupervised clustering methods. Another limitation is that CCST learns the cell embedding using GCNs and performs cell clustering separately. The learned feature is therefore not optimized for the final clustering purpose. An end-to-end workflow is needed to allow the clustering to optimize the embedding parameters in the GCN training. Additionally, considering that, in CCST, single-cell features are only extracted from gene expression information, the model could be further extended to integrate multiple representations, for instance, histological images for morphological features and RNA velocity for cell dynamics. We anticipate that studies in these directions would further improve the performance of the method in the future.

Methods

Datasets. Using technology for imaging the transcriptome in situ with high accuracy, multiple high-throughput spatial expression datasets have become available for analysing cells based on both gene expression and spatial distribution information. Two benchmark datasets are selected. One is MERFISH[5], consisting of the expression of 10,050 genes in 1,368 human osteosarcoma cells from 3 batches (replicates). The second is seqFISH+[35] from mouse OB (cortex), containing 10,000 genes in 2,050 (913) cells assigned to 11 (10) cell types. Additionally, two ST datasets are utilized: the DLPFC and the 10× Visium spatial transcriptomics dataset for human breast cancer. There are 12 samples in the DLPFC dataset, each including up to six cortical layers and white matter. In the annotation of human breast cancer provided by SEDR[32], the tissue is segmented into 20 areas.

Graph construction and data preprocessing. A graph is described by two matrices: an adjacency matrix for representing the graph structure and a feature matrix for representing node attributes. The spatial information about the cells can be represent by an undirected graph, where a cell is represented by a node and an edge connects a pair of cells that are spatially close to each other (see Supplementary Section 2 for details). To balance the weight between the spatial information and the gene expression of an individual cell, we introduce a hyperparameter λ to generate the hybrid adjacency matrix:

$$A = \lambda \times I + (1 - \lambda) A_0, \tag{1}$$

where $I \in R^{N \times N}$ is an identity matrix. For single-cell resolution datasets, CCST assigns cells to different types or subtypes. For datasets without single-cell

resolution, considering that each spot includes multiple cells, CCST aims to separate a whole tissue into different areas or layers. Because the spatial information is supposed to play a more significant role in the area separation, the recommended value of λ is 0.3 for such (ST) datasets but 0.8 for datasets with single-cell resolution (MERFISH and seqFISH+). Supplementary Figs. 20–24 show further validation analysis regarding the choice of the value of λ.

Following studies on these datasets[5,32,35], preprocessing steps are applied to the raw gene count data to extract node features, including filtering out genes with low expression or low variability, normalizing the counts per cell, and batch correction if necessary[32]. Specifically, for the MERFISH dataset with three replicates, we adopt the preprocessing strategy described by the MERFISH study[5], where batch correction is required. After removing genes with low expression of fewer than 1 count per cell on average, we employ Scanorama[49] to correct the batch effect. Next, we normalize the corrected expressions, following equation (2). Finally, genes with low variability as indicated by a variance of the normalized expression lower than 0.4 are dropped[35]. We use

$$\text{expression}_{ij} = \frac{\text{count}_{ij}}{\sum_j \text{count}_{ij}} \times 10,000, \tag{2}$$

where i indexes the cells and j indexes the genes.

For ST and seqFISH+ datasets, similar preprocessing steps are adopted but without batch correction. Following SEDR[32], for the two ST datasets, a gene is filtered out if it is detected by fewer than three spots and, with normalization, the gene expression dimension is reduced to 200 by PCA. For seqFISH+ datasets, before normalization, given their relatively low expression level[35], genes with average expression of less than 0.02 per cell or variance of less than 0.05 across cells are filtered out. Detailed information regarding the final data fed into the GCN model is presented in Supplementary Table 1.

Node embedding and clustering. With the recent progress in GCNs, several approaches to learn node representations from graph-structured data have been proposed. Here, we utilize an unsupervised graph embedding method, DGI[36].

Different from previous approaches based on a random walk, DGI relies on maximizing the mutual information between the local representations and global summaries of graphs. In a GCN, nodes are embedded by repeatedly aggregating the features of neighbouring nodes. The extracted local feature contains the information of a subgraph centred on each individual node. To better explore the high-level feature of the whole graph, DGI is designed to learn an encoder by maximizing the mutual information over patches. This feature contains not only local but also global features.

The input to DGI is the hybrid adjacency matrix $A \in R^{N\times N}$ and a set of node features $X=\{x_1, x_2, \ldots, x_N\}$, where N is the number of nodes, $x_i \in R^F$ represents the features of node i and F is the number of node features. In the standard version of DGI and most applications of GCNs, the adjacency matrix A is assumed to be filled with binary numbers, that is, $A_{ij}=1$ if there exists an edge between node i and j in the graph and $A_{ij}=0$ otherwise. Here, we further apply DGI to the weighted graph constructed with the hybrid adjacency matrix.

The objective of DGI is to learn an encoder E that maps the input feature and adjacency matrix to an embedding space: $E(X, A)=H=\{h_1, h_2, \ldots, h_N\}$, where H represents high-level representations, $h_i \in R^M$ for each node i and M is the number of embedding features. The encoder is composed of four graph convolutional layers for aggregating features over neighbouring nodes with a parametric rectified linear unit as the activation function. The lth graph convolutional layer is

$$H^{(l+1)} = GCN^{(l)}\left(H^{(l)}, A\right) = \sigma\left(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}H^{(l)}W^{(l)}\right) \tag{3}$$

where H^l and H^{l+1} are the input and output of the lth graph convolutional layer and $W^{(l)}$ is the weight matrix used for feature transformation. $\tilde{A}$ is the adjacency matrix after being added by self-loops:

$$\tilde{A} = A + I, \quad I \in R^{N\times N}, \tag{4}$$

$$\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}. \tag{5}$$

The parametric rectified linear unit function is

$$\text{PReLU}(x) = \begin{cases} x, & \text{if } x \geq 0 \\ ax, & \text{otherwise} \end{cases}, \tag{6}$$

where a is a learnable parameter.

The global representation s is obtained by mapping from the local representations with a readout function S: $s=S(\text{E}(X, A))$ and S: $R^{N\times M} \to R^M$. With the local and global features, a discriminator D: $R^M \times R^M \to R$ is introduced to evaluate how much graph/level information is contained in a local patch. A higher $D(h_i, s)$ indicates that the patches are more likely to be contained within the summary. To train the discriminator, we generate negative samples by using a corruption function $C: \bar{A} = C(A)$ where the edges in the graph are reconstructed randomly. We then obtain the local representations $\bar{h}_i$ for negative samples as well. The full objective is

$$L = \sum_{i=1}^{N} E_{X,A}\left[\log D\left(h_i, s\right)\right] + E_{X,\bar{A}}\left[\log\left(1 - D\left(\bar{h}_i, s\right)\right)\right]. \tag{7}$$

By maximizing the approximate representation of the mutual information between h_i and s, DGI outputs a node embedding that contains the structural information of the graph. PCA is performed on the obtained embedding vector for dimension reduction. The k-means++ clustering algorithm is employed on the top principal components to identify cell groups or subpopulations. A validation study regarding the selection of the value of the cluster number k is shown in Supplementary Fig. 25. When applying CCST to a new dataset without knowing the number of desired groups, the silhouette score could be used as a reference for selecting a proper value of k.

Differential gene expression analysis. To verify the different biological functions of each clustered cell subpopulation, we find DE genes that are expressed highly in each subpopulation by using the Mann–Whitney U test for all the cell types. Using the top 200 DE genes with the whole gene list in the corresponding dataset as the background, we carried out GO term enrichment analysis for each subpopulation to construct a functional enrichment profile. According to the GO analysis server (http://geneontology.org/), the statistical significance is obtained according to Fisher's exact test by default, and GO terms with a false discovery rate lower than 0.05 are statistically significant.

Statistics and reproducibility. No statistical methods are used to predetermine the sample sizes, considering that we use all the samples from each dataset. Each of the utilized datasets has been applied in previous study[5,13,31,32]. For the MERFISH dataset, cells in C2 are excluded from further analyses because only two cells are assigned to this cluster. No data were excluded from the analyses on the other datasets. No randomization or blinding is used in this study.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Source data for Figs. 2–6 are available with this manuscript. The datasets utilized in this study can be downloaded from: (1) MERFISH dataset[5]: https://www.pnas.org/doi/10.1073/pnas.1912459116#supplementary-materials or our Github link: https://github.com/xiaoyeye/CCST/tree/main/dataset; (2) SeqFISH+ dataset[35]: https://github.com/CaiGroup/seqFISH-PLUS; (3) DLPFC dataset[37]: http://research.libd.org/globus/jhpce_HumanPilot10x/index.html; (4) 10× Visium spatial transcriptomics data of human breast cancer: https://support.10xgenomics.com/spatial-gene-expression/datasets/1.1.0/V1_Breast_Cancer_Block_A_Section_1. The annotation file can be found on the SEDR[32] website: https://github.com/JinmiaoChenLab/SEDR_analyses/tree/master/data/BRCA1.

Code availability

CCST is implemented in Python. The source code and the utilized MERFISH dataset can be downloaded from the supporting website: https://github.com/xiaoyeye/CCST. https://doi.org/10.5281/zenodo.6560643 (ref. [50]).

Received: 18 October 2021; Accepted: 19 May 2022;
Published online: 27 June 2022

References

1. Codeluppi, S. et al. Spatial organization of the somatosensory cortex revealed by osmFISH. *Nat. Methods* **15**, 932–935 (2018).
2. Moffitt, J. R. & Zhuang, X. RNA imaging with multiplexed error-robust fluorescence in situ hybridization (MERFISH). *Methods Enzymol.* **572**, 1–49 (2016).
3. Moffitt, J. R. et al. High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 11046–11051 (2016).
4. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348**, aaa6090 (2015).
5. Xia, C., Fan, J., Emanuel, G., Hao, J. & Zhuang, X. Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 19490–19499 (2019).
6. Eng, C.-H. L., Shah, S., Thomassie, J. & Cai, L. Profiling the transcriptome with RNA SPOTs. *Nat. Methods* **14**, 1153–1155 (2017).
7. Lubeck, E., Coskun, A. F., Zhiyentayev, T., Ahmad, M. & Cai, L. Single-cell in situ RNA profiling by sequential hybridization. *Nat. Methods* **11**, 360–361 (2014).

8. Wang, X. et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* **361**, aat5691 (2018).
9. Ståhl, P. L. et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353**, 78–82 (2016).
10. Rodriques, S. G. et al. Slide-seq: a scalable technology for measuring genome-wide expression at high spatial resolution. *Science* **363**, 1463–1467 (2019).
11. Stickels, R. R. et al. Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nat. Biotechnol.* **39**, 313–319 (2021).
12. Nichterwitz, S. et al. Laser capture microscopy coupled with Smart-seq2 for precise spatial transcriptomic profiling. *Nat. Commun.* **7**, 12139 (2016).
13. Zhao, E. et al. Spatial transcriptomics at subspot resolution with BayesSpace. *Nat. Biotechnol.* **39**, 1375–1384 (2021).
14. Andersson, A. et al. Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography. *Commun. Biol.* **3**, 565 (2020).
15. Pal, B. et al. Construction of developmental lineage relationships in the mouse mammary gland by single-cell RNA profiling. *Nat. Commun.* **8**, 1627 (2017).
16. Arnol, D., Schapiro, D., Bodenmiller, B., Saez-Rodriguez, J. & Stegle, O. Modeling cell–cell interactions from spatial molecular data with spatial variance component analysis. *Cell Rep.* **29**, 202–211 (2019).
17. Yuan, Y. & Bar-Joseph, Z. GCNG: graph convolutional networks for inferring gene interaction from spatial transcriptomics data. *Genome Biol.* **21**, 300 (2020).
18. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 (2019).
19. Abdelaal, T. et al. A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biol.* **20**, 194 (2019).
20. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
21. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587 (2021).
22. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, P10008 (2008).
23. Traag, V. A., Waltman, L. & Van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).
24. Shekhar, K. et al. Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. *Cell* **166**, 1308–1323 (2016).
25. Pandey, S., Shekhar, K., Regev, A. & Schier, A. F. Comprehensive identification and spatial mapping of habenular neuronal types using single-cell RNA-seq. *Curr. Biol.* **28**, 1052–1065 (2018).
26. Zhu, Q., Shah, S., Dries, R., Cai, L. & Yuan, G.-C. Identification of spatially associated subpopulations by combining scRNAseq and sequential fluorescence in situ hybridization data. *Nat. Biotechnol.* **36**, 1183–1190 (2018).
27. Stoltzfus, C. R. et al. CytoMAP: a spatial analysis toolbox reveals features of myeloid cell organization in lymphoid tissues. *Cell Rep.* **31**, 107523 (2020).
28. Dries, R. et al. Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome Biol.* **22**, 78 (2021).
29. Pham D., et al. stLearn: integrating spatial location, tissue morphology and gene expression to find cell types, cell–cell interactions and spatial trajectories within undissociated tissues. Prerint at *bioRxiv* https://doi.org/10.1101/2020.05.31.125658 (2020).
30. Teng, H., Yuan, Y. & Bar-Joseph, Z. Clustering spatial transcriptomics data. *Bioinformatics* **38**, 997–1004 (2022).
31. Hu, J. et al. SpaGCN: integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nat. Methods* **18**, 1342–1351 (2021).
32. Fu, H., et al. Unsupervised spatial embedded deep representation of spatial transcriptomics. Preprint at *bioRxiv* https://doi.org/10.1101/2021.06.15.448542 (2021).
33. Chen Y., Zhou S., Li M., Zhao F., & Qi J. STEEL enables high-resolution delineation of spatiotemporal transcriptomic data. Preprint at *research square* https://doi.org/10.21203/rs.3.rs-1240258/v1 (2022).
34. Kipf T. N. & Welling M. Semi-supervised classification with graph convolutional networks. In Proc. *International Conference on Learning Representations* (2017). https://openreview.net/forum?id=SJU4ayYgl
35. Eng, C.-H. L. et al. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature* **568**, 235–239 (2019).
36. Veličković P., et al. Deep graph infomax. In Proc. *International Conference on Learning Representations* (2019). https://openreview.net/forum?id=rklz9iAcKQ
37. Maynard, K. R. et al. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nat. Neurosci.* **24**, 425–436 (2021).
38. Donjerkovic, D. & Scott, D. W. Regulation of the G1 phase of the mammalian cell cycle. *Cell Res.* **284**, C349–364 (2000).
39. Tripathi, V. et al. Long noncoding RNA MALAT1 controls cell cycle progression by regulating the expression of oncogenic transcription factor B-MYB. *PLoS Genet.* **9**, e1003368 (2013).
40. Wang, J. et al. MALAT1 promotes cell proliferation in gastric cancer by recruiting SF2/ASF. *Biomed. Pharmacother.* **68**, 557–564 (2014).
41. Merlot, S., Gosti, F., Guerrier, D., Vavasseur, A. & Giraudat, J. The ABI1 and ABI2 protein phosphatases 2C act in a negative feedback regulatory loop of the abscisic acid signalling pathway. *Plant J.* **25**, 295–303 (2001).
42. Mahdessian, D. et al. Spatiotemporal dissection of the cell cycle with single-cell proteogenomics. *Nature* **590**, 649–654 (2021).
43. Sakaue-Sawano, A. et al. Visualizing spatiotemporal dynamics of multicellular cell-cycle progression. *Cell* **132**, 487–498 (2008).
44. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
45. Cheng, C. et al. Cloning, expression and characterization of a novel human VMP gene. *Mol. Biol. Rep.* **29**, 281–286 (2002).
46. Li, S. et al. Endothelial cell-derived GABA signaling modulates neuronal migration and postnatal behavior. *Cell Res.* **28**, 221–248 (2018).
47. Russ, A. P. et al. Eomesodermin is required for mouse trophoblast development and mesoderm formation. *Nature* **404**, 95–99 (2000).
48. Taberner, L., Bañón, A. & Alsina, B. Sensory neuroblast quiescence depends on vascular cytoneme contacts and sensory neuronal differentiation requires initiation of blood flow. *Cell Rep.* **32**, 107903 (2020).
49. Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat. Biotechnol.* **37**, 685–691 (2019).
50. Li J., Chen S., Pan X., Yuan Y., & Shen H.-B. Cell clustering for spatial transcriptomics data with graph neural networks. *Zenodo* https://doi.org/10.5281/zenodo.6560643 (2022).

Acknowledgements

This work was supported by grants from the National Natural Science Foundation of China (no. 61725302 to H.S.), 62073219 (to H.S.), 62103262 (to Y.Y.) and 61903248 (to X.P.) and the Shanghai Pujiang Programme (no. 21PJ1407700 to Y.Y.).

Author contributions

H.S. and Y.Y. conceived and supervised the study. Y.Y. designed experiments. J.L. developed the computational model and conducted data analysis. Y.Y., H.S. and X.P. provided advice on data analysis. Y.Y. and S.C. proposed the proper computational model. J.L. drafted the manuscript. Y.Y. and H.S. revised the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43588-022-00266-5.

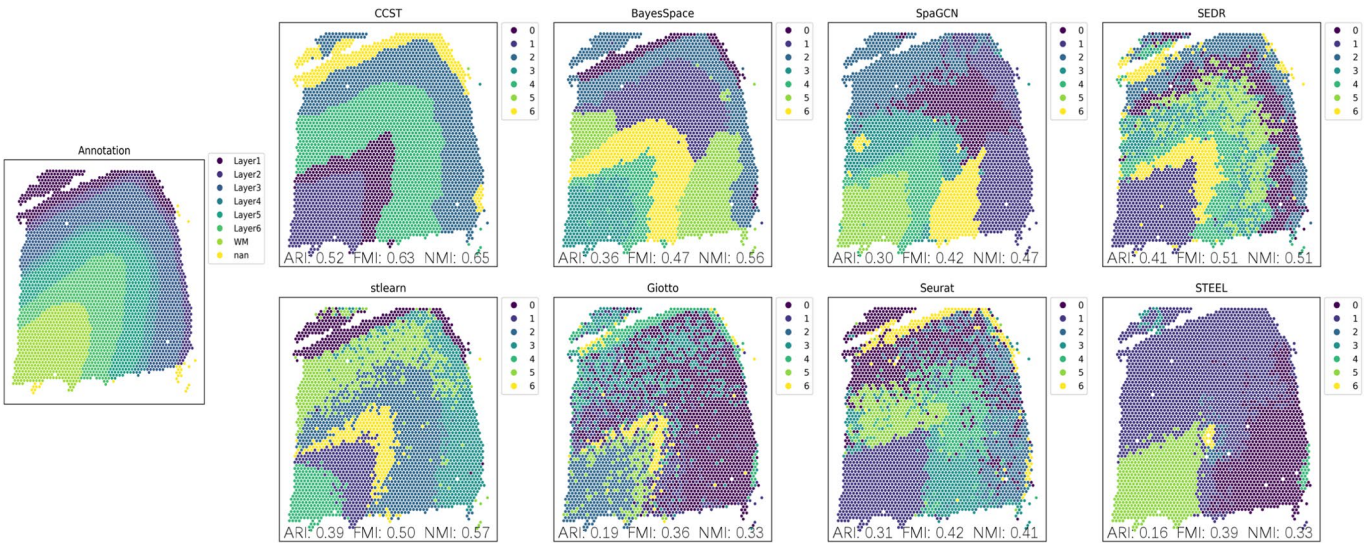
Correspondence and requests for materials should be addressed to Ye Yuan or Hong-Bin Shen.

Peer review information *Nature Computational Science* thanks Xin Zhou and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Handling editor: Kaitlin McCardle, in collaboration with the *Nature Computational Science* team.

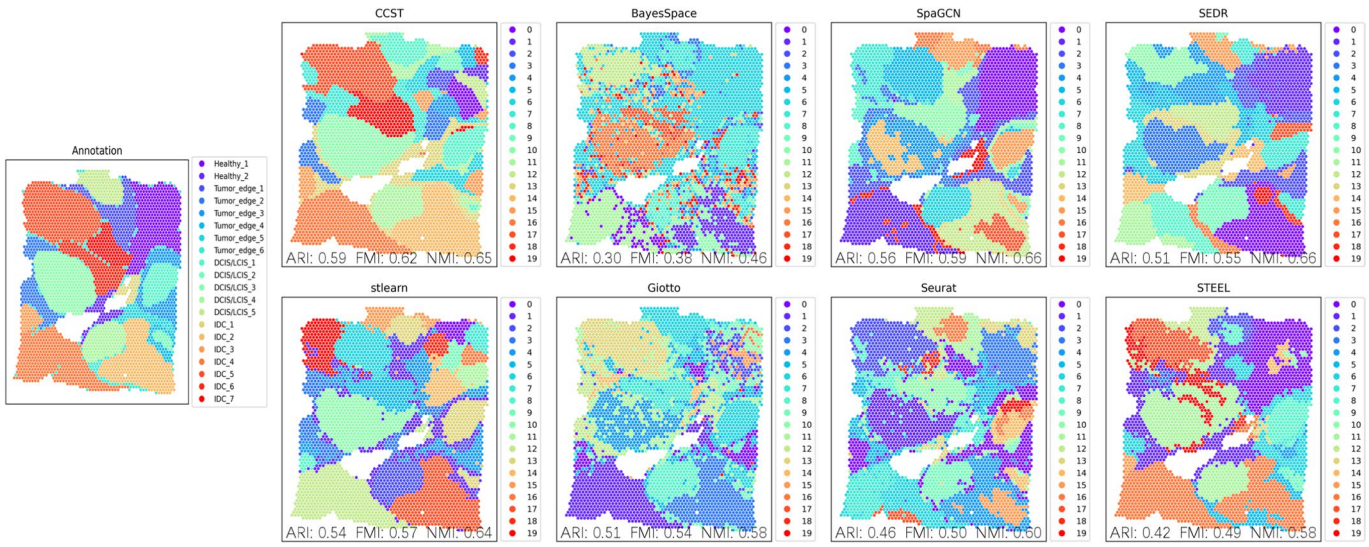
Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2022



Extended Data Fig. 1 | Comparison on sample 151676 of DLPFC. Annotation and cluster labels obtained by CCST and prior methods on sample 151676 of DLPFC. Metrics including ARI, FMI and NMI are annotated on the bottom of each figure. Numbers in the legend refer to cluster labels.



Extended Data Fig. 2 | Comparison on 10x Visium spatial transcriptomics data of human breast cancer. Annotation and cluster labels obtained by CCST and prior methods on 10x Visium spatial transcriptomics data of human breast cancer. Metrics including ARI, FMI and NMI are annotated on the bottom of each figure. Numbers in the legend refer to cluster labels.

nature research

Corresponding author(s): Ye Yuan, Hong-Bin Shen

Last updated by author(s): May 17, 2022

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software is used.
Data analysis	CCST is implemented in Python. The source code and the utilized MERFISH dataset can be downloaded from the supporting website, https://github.com/xiaoyeye/CCST. DOI: 10.5281/zenodo.6560643

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Source data for Figures 2-6 are available with this manuscript.
Datasets utilized in this study can be downloaded from
(1) MERFISH dataset:
https://www.pnas.org/doi/10.1073/pnas.1912459116#supplementary-materials.
We also uploaded it to our Github link:
https://github.com/xiaoyeye/CCST/tree/main/merfish.
(2) SeqFISH+ dataset:

https://github.com/CaiGroup/seqFISH-PLUS
(3) DLPFC:
http://research.libd.org/globus/jhpce_HumanPilot10x/index.html.
(4) 10x Visium spatial transcriptomics data of human breast cancer:
https://support.10xgenomics.com/spatial-gene-expression/datasets/1.1.0/V1_Breast_Cancer_Block_A_Section_1.
The annotation file can be found from:
https://github.com/JinmiaoChenLab/SEDR_analyses/tree/master/data/BRCA1.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	CCST is tested on four datasets, including two Spatial Transcripts datasets, one MERFISH dataset and one SeqFISH+ dataset. Each of these datasets has been utilized in previous study (SpaGCN [1], BayesSpace [2], MERFISH study [3] etc.), which demonstrate that these datasets are believed to be with proper sample size. 1. Hu J, et al. (2021) SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. Nature Methods:1-10. 2. Zhao E, et al. (2021) Spatial transcriptomics at subspot resolution with BayesSpace. Nature Biotechnology:1-10. 3. Xia C, Fan J, Emanuel G, Hao J, & Zhuang X (2019) Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression. Proceedings of the National Academy of Sciences 116(39):19490-19499.
Data exclusions	During the data preprocessing of CCST, low expression genes are filtered out. This preprocessing strategy is widely applied in various gene analysis methods.
Replication	The code is provided for replication purpose. The results can be reproduced by running the code with default parameters. Considering that CCST is a deep learning based model, minor difference may appear during the replication.
Randomization	This is no relevant to our study. Because CCST is a cluster method that can deal with the whole dataset without the requirement of data allocation.
Blinding	This is no relevant to our study. Because CCST is a cluster method that can deal with the whole dataset without the requirement of data allocation.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging